Universidad Carlos III de Madrid

Escuela Politécnica Superior

Departamento: Mecánica de Medios Continuos y Teoría de Estructuras

Tutores ENSAM de París: Wafa Skalli y Alexandre Templier

Tutor UC3M: Dr. Ángel Arias Hernández

INGENIERÍA INDUSTRIAL

RESUMEN DEL PROYECTO FIN DE CARRERA

MODELOS PREDICTIVOS DE COMPORTAMIENTO CON APLICACIÓN BIOMECÁNICA

Abel Mariana Muñoz

Septiembre, 2010

ÍNDICE

CAPÍTULO 1. INTRODUCCIÓN

La escoliosis idiopática es una deformación tridimensional de la columna vertebral cuyo origen aún se desconoce. Se trata de una patología de origen multifactorial. El seguimiento médico de tal patología es necesario para identificar y prevenir toda agravación seria que pudiera conducir a problemas importantes a nivel de sistemas vitales del paciente.

Esta deformación aparece generalmente en preadolescentes en edad de crecimiento. La identificación precoz de los pacientes que presentan riesgo real de agravación es necesaria para una solución terapéutica eficaz. Sin embargo, el diagnóstico es difícil con los medios de análisis actuales.

Si la escoliosis se agrava, los medios de corrección ortopédicos consisten en escayolas o corsé. En caso de que estos medios fracasasen, se aplicarían medios quirúrgicos.

Por lo tanto, para los médicos es esencial disponer de técnicas y medios de seguimiento y diagnóstico eficaces que permitan identificar los factores de riesgo de agravación de la escoliosis idiopática.

Se puede concluir que la actualmente **la problemática** es que los médicos no pueden determinar, en un primer examen clínico, si la escoliosis se agravará o no.

CAPÍTULO 2.TRABAJOS PRECEDENTES

El proyecto forma parte de un gran proyecto llevado a cabo en el departamento de biomecánica de la ENSAM de París.

Anteriormente a este proyecto se realizó la **Tesis de Nicolas Champain : "** Investigación de los factores biomecánicos en la agravación de la escoliosis idiopática".

Esta tesis consiste principalmente en un trabajo de investigación en el que una base de datos ha sido creada gracias al seguimiento médico de 72 pacientes que padecían escoliosis idiopática moderada. Los información de los pacientes ha sido recolectada de varias clínicas de Francia.

El conjunto de datos medidos y registrados para cada paciente son de varios tipos: datos generales, exámenes clínicos, radiografías y datos posturales y de movilidad.

Todos estos parámetros no pueden ser utilizados conjuntamente para efectuar test estadísticos, por lo que se han seleccionado algunos de estos parámetros con el fin de reducir el número de variables por paciente y para que los datos sean los más homogéneos posibles para el estudio estadístico.

Con la ayuda de expertos médicos, 6 parámetros han sido elegidos: La rotación axial de la vértebra apical, la rotación intervertebral de la zona de unión superior, la rotación intervertebral de la zona de unión inferior, el índice de hipo cifosis apical, el índice de torsión y el ángulo de Cobb.

Actualmente en el laboratorio se trabaja con estos 6 parámetros tomados en un primer examen clínico del paciente, usando la técnica del **análisis factorial discriminante**. Está técnica se utiliza de la siguiente forma; teniendo en el laboratorio una base de datos de 27 pacientes que padecen escoliosis idiopática moderada, se comparan estos 6 parámetros en los 27 pacientes y estos mismos 6 parámetros en sujetos sanos y en sujetos con escoliosis severa, con el fin de encontrar semejanzas y poder equivalencia determinar si la escoliosis se agravará o, por el contrario, permanecerá estable.

Con el método del análisis factorial discriminante hemos obtenido los siguientes resultados:

	Devenir real de los pacientes		
Resultados de la prueba	Similar a la escoliosis severa	Similar a sujeto sano	No clasificada
Agravamiento N=17	**12 (71 %)**	3 (17%)	**2 (12 %)**
Estable N=10	0 (0 %)	**10** (100 %)	**0**

Tabla 1.Resultados obtenidos con el análisis factorial discriminante

Se ha obtenido un total de 81% de sujetos bien clasificados

CAPÍTULO 3. OBJETIVO DEL PROYECTO

El objetivo del proyecto es proponer soluciones alternativas a los métodos estadísticos de predicción actuales para mejorar la evaluación del riesgo de agravación de la escoliosis idiopática en un primer examen. Estas soluciones deben determinar si la escoliosis evolucionará hacía escoliosis más severas y cuáles son los parámetros más importantes que la hacen agravarse.

Para alcanzar este objetivo, el proyecto se ha dividido en dos partes:

- En la primera parte se ha realizado una investigación bibliográfica de los métodos actuales de "data mining" y se han seleccionado 4 técnicas que pueden ajustarse a nuestras necesidades.
- En la segunda parte se han utilizado las 4 técnicas seleccionadas anteriormente y se han obtenido resultados. Además, se obtendrán resultados mediante el algoritmo Q-Finder, aportado por la empresa Quinten[1].

[1] http://quinten-france.com/EN/indexEN.html

CAPÍTULO 4. INVESTIGACIÓN BIBLIOGRÁFICA

En este capítulo primeramente se explica cómo funcionan las técnicas de data mining. A continuación se exponen las 4 técnicas escogidas tras haber realizado la investigación bibliográfica.

4.1 Data mining

Data mining o minería de datos es el proceso de extracción de conocimiento válido, útil y comprensible a partir de datos recolectados [Witten & Frank 2000].

En las técnicas de data mining hay dos fases. Una fase de aprendizaje para elaborar un modelo que sintetizará las relaciones entre las variables y a continuación una fase deductiva que podrá ser aplicada a un nuevo conjunto de datos con el fin de clasificar o predecir un valor. A continuación se describen brevemente estas etapas:

1. Aprendizaje: Consiste en la construcción del modelo sobre una primera muestra para la que se conocen los valores de la variable a predecir.
2. Test: Verificación del modelo sobre una segunda muestra para la que también se conocen los valores de las variables a predecir. El objetivo es comparar los resulados obtenidos con el modelo y los valores reales. Si el resultado del test no es satisfactorio volvemos a realizar la fase de aprendizaje.
3. Validación del modelo sobre una tercera muestra, para tener una idea de la tasa de error.
4. Aplicación del modelo en un conjunto de datos de la población a estudiar para determinar el valor de la variable objetivo.

Validación del modelo

¿Por qué hay que validar el modelo?

Hay que verificar si el modelo construido se ajusta a las necesidades. Los modelos pueden dar falsos resultados, ser poco eficaces o estar demasiado ajustados al conjunto de datos utilizados en la fase de aprendizaje, es decir, que la aplicación del modelo a futuros conjuntos de datos sería poco eficaz estar poco generalizado.

Herramientas para la validación del modelo

Para determinar la validez del modelo utilizaremos dos parámetros: sensibilidad y especificidad.

Sensibilidad: Es la probabilidad de clasificar bien un paciente con la enfermedad, es decir, la probabilidad de obtener un resultado positivo en la prueba cuando el sujeto padece realmente la enfermedad. La sensibilidad es, por tanto, la capacidad de detectar la enfermedad.

Especificidad: Es la probabilidad de clasificar bien un paciente sano, es decir, la probabilidad de obtener un resultado negativo en la prueba cuando el sujeto no padece la enfermedad. La especificidad es, por tanto, la capacidad de detectar los sujetos sanos.

Si colocamos los resultados de los pacientes en una tabla (llamada **matriz de confusión)** es posible calcular la sensibilidad y la especificidad.

Resultados de la prueba	**Diagnóstico real**	
	Enfermo	**Sano**
Positivo	Verdadero positivo (VP)	Falso positivo (FP)
Negativo	Falso negativo (FN)	Verdadero negativo (VN)

Tabla 2.Matriz de confusión

$$Sensibilidad = \frac{VP}{VP+FN} \qquad (4.1)$$

$$Especificidad = \frac{VN}{VN+FP} \qquad (4.2)$$

Si detectamos todos los sujetos como positivos = Sensibilidad 100% y Especificidad 0%

Si no detectamos ningún sujeto como positivo = Sensibilidad 0% y Especificidad 100%

El objetivo de la validación del modelo es obtener una sensibilidad del 100% y una especificidad del 100%

4.2 Métodos estadísticos

Después de hacer la documentación sobre los métodos de data mining actuales, estos se han agrupado en 4 grandes grupos: árboles de decisión, redes neuronales, clustering y ensemble learning, de los cuáles los tres primeros serán utilizados para hacer la predicción y el método ensemble learning será utilizado para encontrar reglas de asociación entre parámetros.

Además de estos cuatro métodos, se utilizará el algoritmo Q-FINDER, de la fundación Quinten[2], que será utilizado también para encontrar reglas de asociación.

[2] http://quinten-france.com/EN/indexEN.html

4.2.1 ÁRBOLES DE DECISIÓN

A continuación se explica el funcionamiento de los árboles de decisión.

- Principio

Para cada atributo y recursivamente, el algoritmo divide los datos y selecciona la division que aporte mayor información

- Descripción

Los árboles de decisión se utilizan para determinar el valor combinado de una serie de acciones que ocurren sucesivamente de acuerdo a unas probabilidades. El objetivo es determinar un valor final en función de unas condiciones con el fin de poder tomar una decisión.

- Algoritmo: **C4.5**
- Funcionamiento del algoritmo

El algoritmo C4.5 genera un árbol de decisión, a partir de los datos, gracias a divisiones realizadas recursivamente.

El algoritmo considera todos los ensayos posibles que puedan dividir el conjunto de datos y selecciona el resultado que aporte la mayor infomación.

Para cada atributo discreto, el algoritmo considera un ensayo con *n* resultados, donde *n* es el número de valores posibles del atributo.

Para cada atributo continuo, el algoritmo realiza un ensayo binario. Si una variable A posee valores numéricos continuos, se realiza una prueba binaria con resultados del tipo $A \leq Z$ y $A > Z$.

La salida es una árbol que muestra las variables más discriminantes elegidas por el algoritmos, así como los valores límites que definen cada rama del árbol.

La figura siguiente muestra un ejemplo de un árbol de decisión utilizado en el departamento de finanzas de una empresa para decidir sobre una nueva posible inversión.

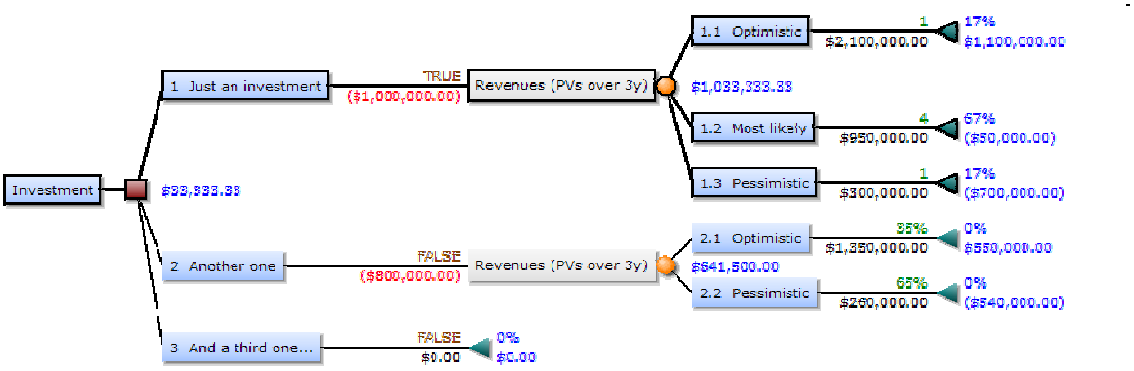


Figura 1. Ejemplo de árbol de decisión

4.2.2 REDES NEURONALES

A continuación se explica el funcionamiento de las redes neuronales.

- Principio

Se modifican la importancia de los parámetros y los enlaces entre las neuronas gracias al aprendizaje con la experiencia.

- Descripción

Las redes neuronales están formadas por unidades de tratamiento que intercambian información. Se utilizan para reconocier patrones. Tienen la capacidad de aprender y mejorar su funcionamiento. Son capaces de aprender con la experiencia y de extraer características esenciales a partir de datos a priori no importantes.

Durante el proceso de aprendizaje los pesos de los enlaces entre las neuronas se ajustan para obtener un resultado específico. Una red neuronal no necesita un algoritmo para alcanzar un resultado final puesto que lo obtiene gracias a la modificación de los pesos de los enlaces. Sin embargo, hace falta un buen **algoritmo de aprendizaje** que dé a la red la capacidad de discriminar gracias a un entrenamiento con patrones.

- Algorimo: **PERCEPTRON**
- Funcionamiento del algorimo

Cada neurona *i* está caracterizada por una función de activación que transforma los valores de entrada en una señal de salida. Esta señar es enviada a través de los canales de comunicación a las otras unidades de la red. En estos canales la señal es modificada conforme a los pesos asociados a cada canal.

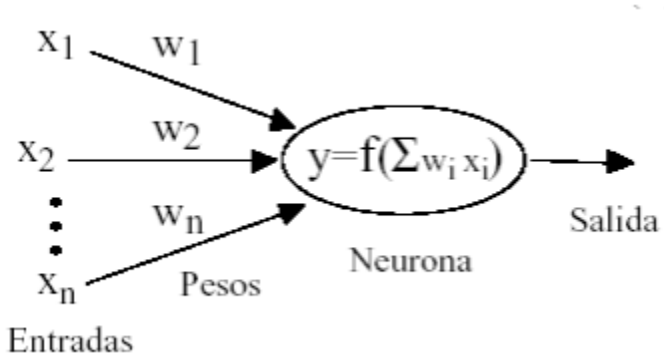


Figura 2. Funcionamiento de una neurona

Todas las neuronas tienen un estado. Existen dos estados posibles: reposo y activada.

Se considera y_i como el valor de salida de una neurona en un instante concreto i. Una neurona recibe un conjunto de señales que le aportan información sobre el estado de activación de las otras neuronas con las que está conectada. La salida de la última neurona es la suma del valor de cada neurona multiplicada por el valor del peso de la únion.

$$net_j = \sum_i^N w_{ji} * y_i \quad (4.3)$$

El PERCEPTRON utiliza un aprendizaje supervisado y un algoritmo de aprendizaje por corrección de error. El aprendizaje supervisado consiste en un tipo de entrenamiento donde se le proporciona al sistema información sobre las entradas y salidas. De esta manera, el sistema tiene un punto de referencia para evaluar el funcionamiento basándose en la diferencia entre estos valores y poder modificar los parámetros de la red.

Finalmente, la red encontrará la función más próxima posible a la función óptima.

La salida es un valor entre -1 y 1 que es función de todos los valores de las neuronas y de todos los pesos de las uniones de la red.

La figura siguiente muestra la estructura de una red neuronal.

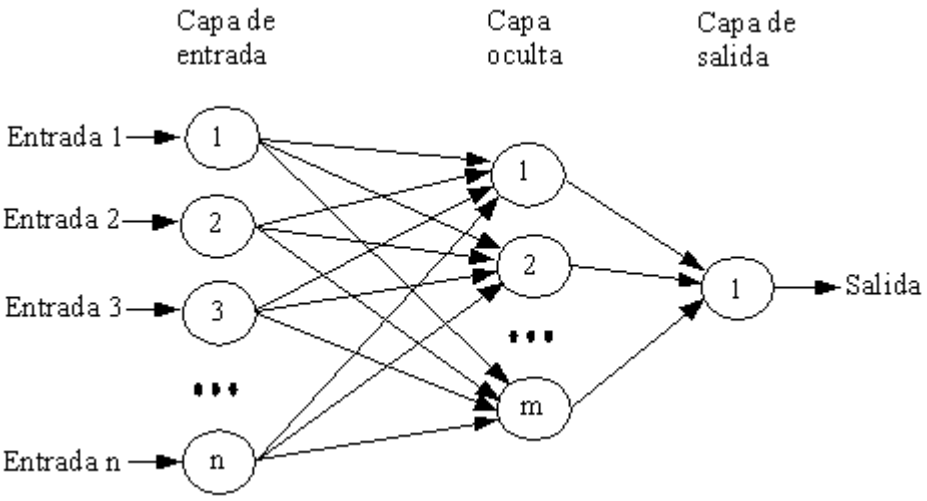


Figura 3. Estructura de una red neuronal

4.2.3 ENSEMBLE LEARNING

A continuación se explica el funcionamiento de las redes neuronales.

- Principio

Combinación de funciones de predicción gracias a un algoritmo de aprendizaje

- Algoritmo: **RULEFIT**[3]
- Descripción

Normalmente, el proceso a seguir para la aplicación de un método estadístico es elegir el método con el error más pequeño posible, pero eso no es suficiente. Lo que se pretende conseguir con un algoritmo del tipo ensemble learning es eliminar los riesgos de elegir solamente una función de predicción.

El objetivo es predecir el valor de una variable utilizando varias funciones que son combinadas para obtener un resultado único. Se trata de una regresión lineal en la que cada variable independiente es función de predicción (basada en el conjunto de datos inicial). La función de salida es una combinación lineal de las funciones de predicción.

Ejemplo:

Los resultados que salen del algoritmo son mostrados en una tabla en la que se puede ver el número de regla, las variables que pertenecen a la regla y los valores mínimos y máximos de cada variable.

[3] http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf Jerome H.Friedman et Bogdan E. Popescu. 5 octobre 2005

Regla 1	Variables	Valor mínimo	Valor máximo
	Incidencia pelviana	-∞	56.35
	RIA_Vertebra_Sup	-2.65	∞

Tabla 3. Modo de mostrar los resultados con la técnica Ensemble learning

La predicción del algoritmo es una combinación lineal de las reglas resultantes.

4.2.4 CLUSTERING

A continuación se explica el funcionamiento del clustering

- Principio

Clasificación de datos gracias a las semejanzas y diferencias que hay entre ellos.

- Descripción

Los algoritmos de tipo clustering permiten clasificar un conjunto de datos en un número determinado de grupos.

- Algoritmo: **K-MEANS**
- Funcionamiento del algoritmo

Los métodos de clasificación, también llamados análisis en cluster, permiten dividir un juego de datos en subgrupos homogéneos. Una característica importante es que los clustesr no son definidos a partir de una variable externa sino a partir de las propia estructura de los datos.

El número de clusters es fijado previamente. La idea principal es definir *k* centros, uno por cluster. Estos centros son situados con la mayor distancia posible entre ellos.

El paso siguiente es tomar cada punto que pertenece a la base de datos y asociarlo al centro más próximo. Cuando no quedan más puntos por asociar, el primer paso está completado y el primer agrupamiento queda definido.

A continuación se recalcula un nuevo *k* como baricentro de los clusters resultantes del primer paso. Después de obtener estos nuevos centros, se crea un bucle. Como resultado de cada bucle los centros cambian su localización paso a paso hasta que los clusters dejan de cambiar su situación.

El objetivo es minimizar el error cuadrático. La función objetivo es:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (4.4)$$

Donde $||xi(j)-cj||^2$ es la distancia medida entre el punto xi(j) y el centro del cluster.

La salida es una tabla con los clusters donde cada muestra ha sido clasificada así como el número de muestras de cada cluster.

4.2.5 Q-FINDER

A continuación se explica el funcionamiento del algoritmo Q-finder.

- Principio

El algoritmo identifica las combinaciones de características asociadas a una densidad del fenómeno de interés superior a la densidad media observada.

- Funcionamiento

En el conjunto de datos, el algoritmo encuentra grupos de pacientes donde la densidad de las escoliosis que se agravan es superior a la densidad observada sobre toda la población. Estos grupos son caracterizados por un conjunto de reglas.

Para definir cada grupo, Q-finder utiliza las variables de la base de datos, limitando cada una entre dos valores.

La combinación de estas variables y sus valores máximos y mínimos delimita el conjunto de pacientes que definen la regla.

Ejemplo:

La salida del Q-finder se muestra de la siguiente forma:

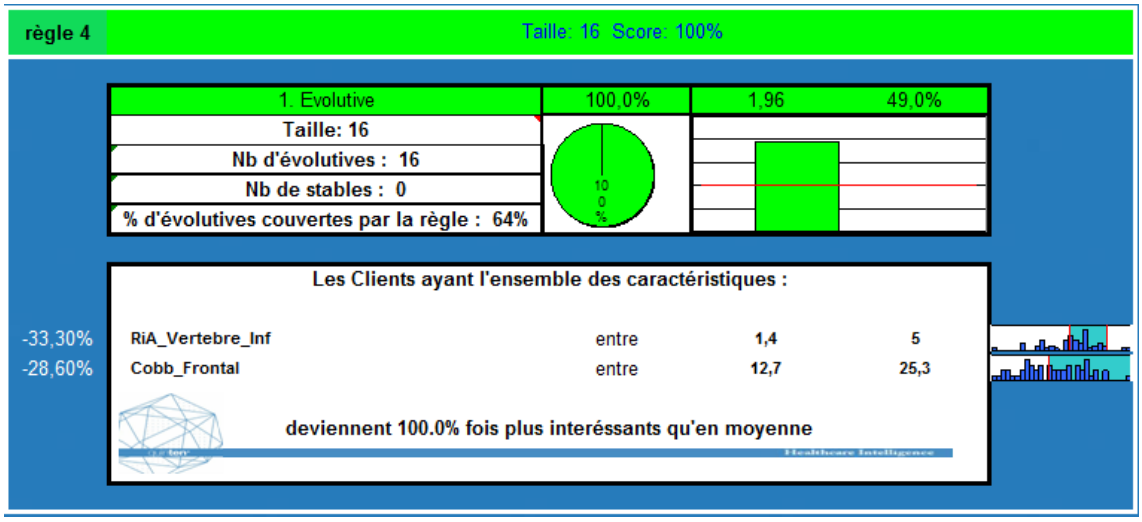


Figura 4. Ejemplo del algoritmo Q-finder

Las reglas son mostradas por orden de importancia. En cada regla se pueden ver, además del número de regla y sus variables, el número de individuos que la regla cubre y el porcentaje de escoliosis evolutivas así como otros parámetros.

Además podemos añadir o quitar variables, modificar los valores máximos y mínimos o combinar dos o más reglas.

CAPÍTULO 5. METODOLOGÍA PARA LA OBTENCIÓN Y ANÁLISIS DE RESULTADOS

En esta sección se presentan los resultados obtenidos con cada algoritmo. Primeramente, se presentarán los datos sobre los que se va a trabajar. Se utilizará un conjunto de datos que está compuesto por 49 pacientes con escoliosis moderada (24 estables y 25 evolutivas). Todos los pacientes son adolescentes que han sido examinados en la Unidad de Raquis del hospital Bellevue en St. Etienne por los doctores I. Courtois y E. Ebermeyer.

Para cada paciente se ha realizado un examen radiográfico y una reconstrucción 3D en su primer examen clínico. Los parámetros resultantes de esta reconstrucción se muestran en la siguiente tabla:

La cifosis T4/T12	La lordosis L1/L5	La pente sacrée
La versión pelviana	La incidencia pelviana	La inclinación media de la columna en el plano frontal
La inclinación media de la columna en el plano sagital	La rotación vertebra axial de la vertebra apical	La rotación intervertebral de la zona de unión superior
La rotación intervertebral de la zona de unión inferior.	El índice de hipocifosis apical	El índice de torsión
	El ángulo de Cobb	

Tabla 4. Parámetros resultantes de la reconstrucción

Se utilizarán dos programas para lanzar los algoritmos: R[4] y WEKA[5], que son dos programas de software libre que pueden descargarse gratuitamente de sus respectivas páginas web. Para el C4.5, el PERCEPTRON, y el K-MEANS se usará WEKA y para el RULEFIT el programa R.

Además, para el análisis del algoritmo Q-finder utilizaremos un programa creado sobre Microsoft Excel por la fundación Quinten[6].

Se utilizarán 3 métodos para la evaluación:

Autovalidación: Mismo conjunto de datos para el aprendizaje y para la validación.

[4] http://www.r-project.org/
[5] http://www.cs.waikato.ac.nz/ml/weka/
[6] http://quinten-france.com/EN/indexEN.html

Leave one out: Es un tipo de técnica dentro del método cross validation.

Cross-validation-k: Dado un número *k* se hace la división de los datos en *k* partes y, para cada una de las partes, se hace el aprendizaje con las *k-1* partes restantes y validamos con la parte elegida. Leave one out es como lanzar el algoritmo con cross validation k donde *k* es igual al número de pacientes. De esta manera cogemos un individuo del conjunto para hacer la validación y el resto para el aprendizaje.

Ejemplo:

Si, por ejemplo, tenemos 49 pacientes, hacemos "cross validation k" con *k=49*, de esta manera hacemos la validación con un paciente y el aprendizaje con los 48 restantes.

Validación con 10 pacientes: A partir del conjunto de datos, se han escogido aleatoriamente 10 pacientes (5 estables y 5 evolutivos). Se usarán estos 10 pacientes para validar el modelo y el aprendizaje se realizará con los 39 restantes.

Utilizaremos las 3 técnicas anteriores para los algoritmos C4.5, PERCEPTRON, K-MEANS y RULEFIT. Para analizar los resultados utilizaremos matrices de confusión así como la sensibilidad y la especificidad.

Los resultados extraídos del Q-finder son aportados por la fundación QUINTEN. Los parámetros utilizados para este algoritmo son todos los parámetros arriba comentados además de varios parámetros clínicos.

Utilizaremos los resultados obtenidos para extraer información y poder evaluar la utilidad de cada método.

CAPÍTULOS 6. RESULTADOS

6.1 Resultados C4.5 (Árboles de decisión)

A continuación les resultados salidos de lanzar el algoritmo C4.5

- Autovalidación

Matriz de confusión

a	b	←clasificado como
22	2	**a = Estable**
0	25	**b = Evolutiva**

Tabla 5. Matriz de confusión para la autovalidación con el C4.5

$$Sensibilidad(\%) = \frac{25}{25+0} = 100\%$$

$$Especificidad(\%) = \frac{22}{22+2} = 91.7\%$$

Hemos obtenido un 95.92% de sujetos bien clasificados y el árbol resultante es el siguiente:

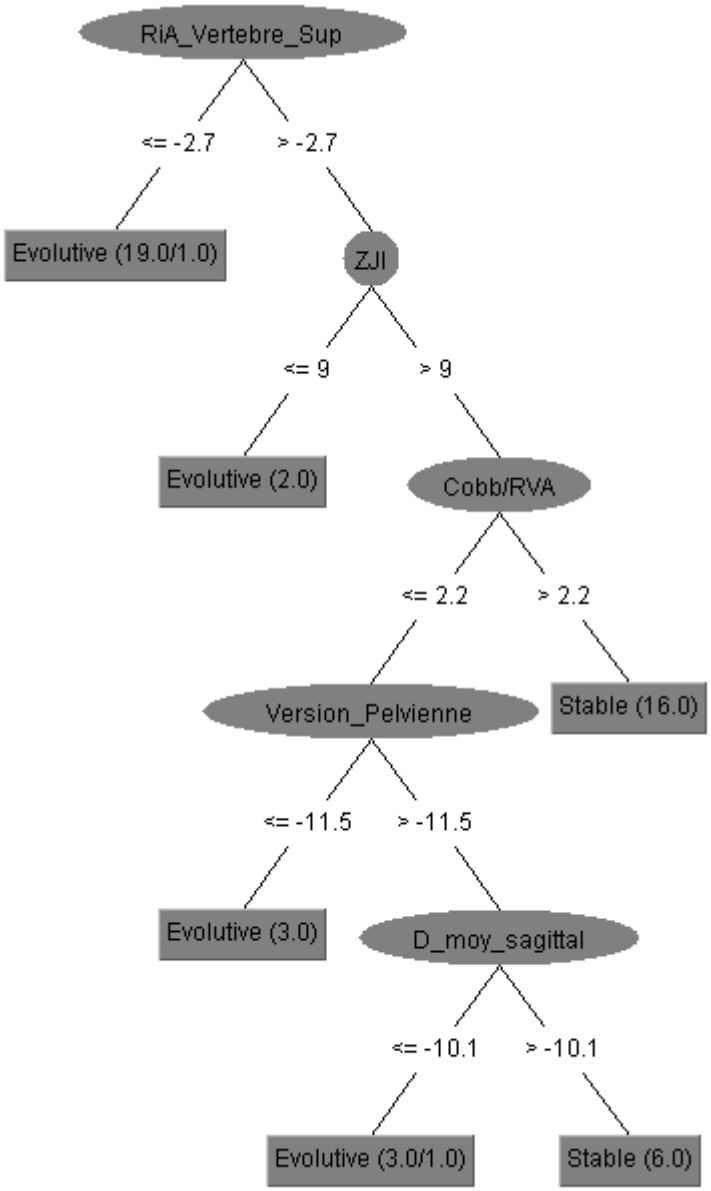


Figura 5.Árbol para la autovalidación

- Leave one out

Con esta técnica hemos obtenido los siguientes resultados:

Matriz de confusión

a	**b**	**←clasificado como**
16	8	**a = Estable**
8	17	**b = Evolutiva**

Tabla 6. Matriz de confusión para leave one out con el C4.5

$$Sensibilidad(\%) = \frac{17}{17+8} = 68\%$$

$$Especificidad(\%) = \frac{16}{16+8}$$

Hemos obtenido un 67.35% de pacientes bien clasificados y el algoritmo nos ha dado el siguiente árbol:

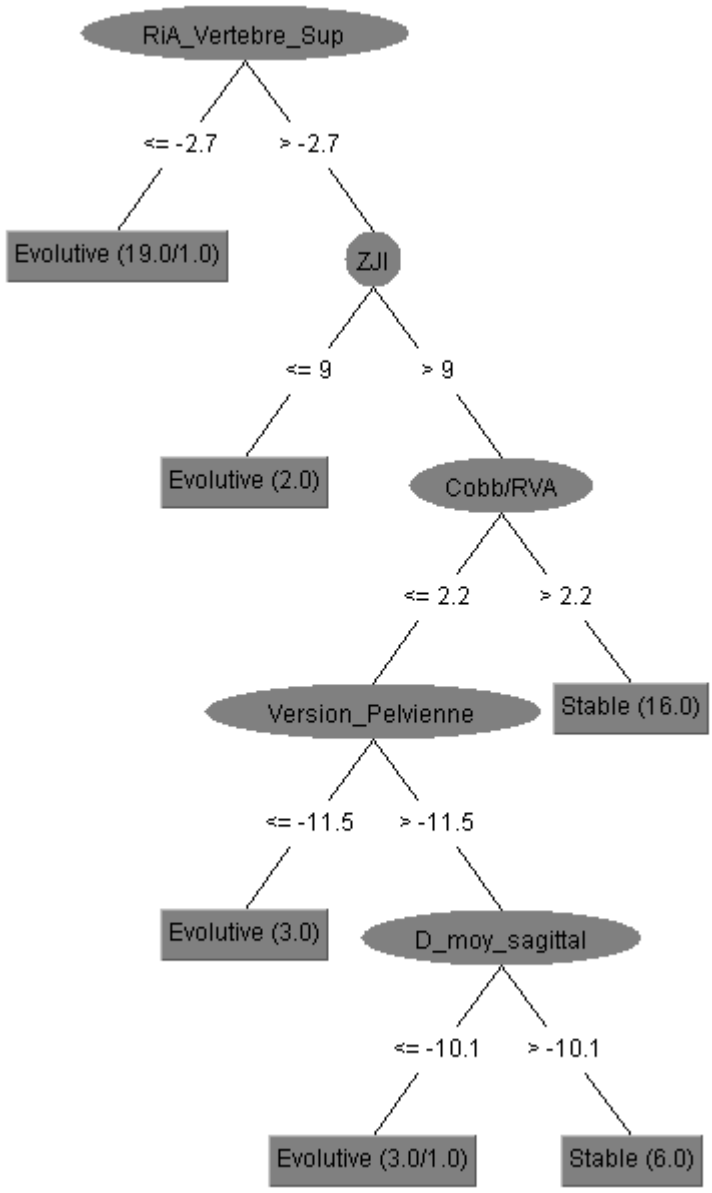


Figura 6.Árbol para el "leave one out"

- Validación con 10 pacientes

Con esta técnica se han obtenido los siguientes resultados:

Matriz de confusión

a	b	←clasificado como
2	3	**a = Estable**
3	2	**b = Evolutiva**

Tabla 7. Matriz de confusión para validación con 10 pacientes con el C4.5

$$Sensibilidad(\%) = \frac{2}{2+3} = 40\%$$

$$Especificidad(\%) = \frac{2}{2+3} = 40\%$$

Hemos obtenido un 40 % de pacientes bien clasificados y el algoritmo nos ha dado el siguiente árbol:

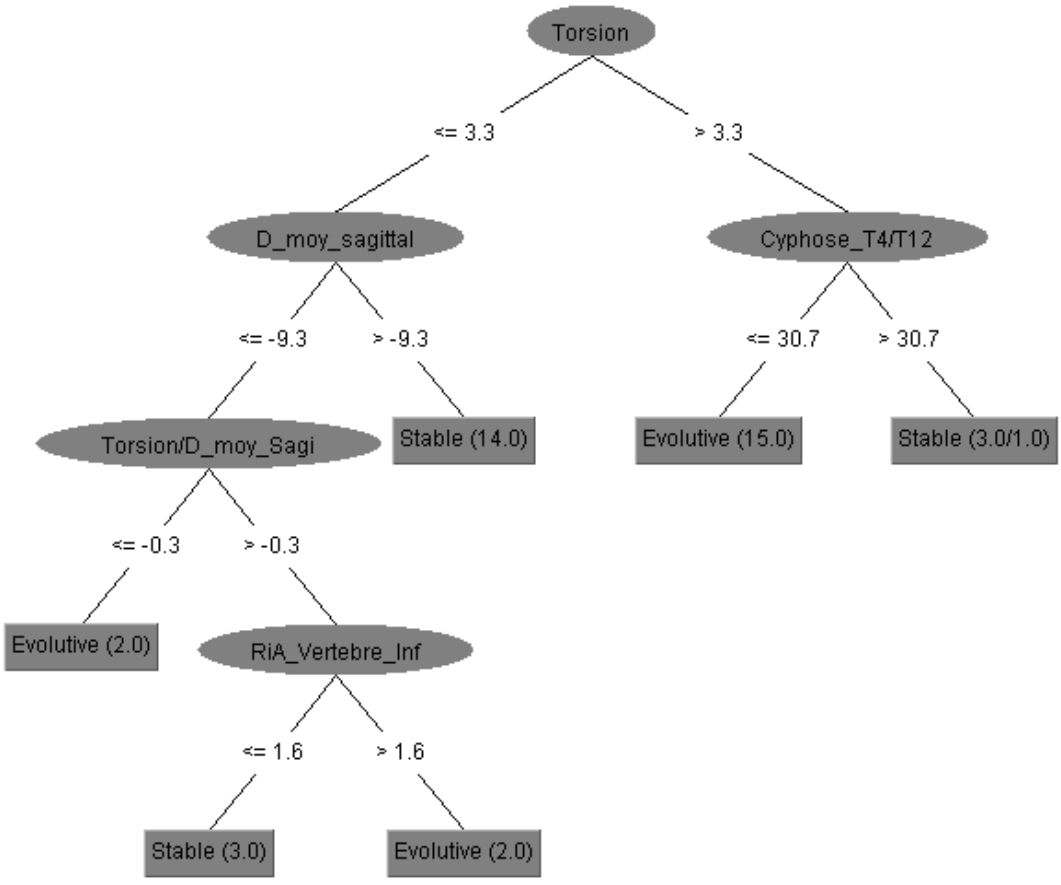


Figura 7.Árbol con validación con 10 pacientes

6.1.1 CONCLUSIONES DE LOS RESULTADOS CON EL C4.5

Hemos obtenido un porcentaje de individuos bien clasificados más grande con la autovalidación. Hay que prestar atención a este resultado porque con la autovalidación los resultados serán siempre optimistas puesto que validamos sobre el mismo conjunto de datos sobre los que hemos hecho el aprendizaje.

Leave one out y la validación elaboran el mismo árbol de decisión pero el árbol del leave one out corresponde a un árbol medio. La disminución de los buenos resultados con leave one out muestra una débil robustez y gran influencia del conjunto de datos del aprendizaje.

Se puede ver que cuando hacemos la validación sobre un grupo diferente del grupo de aprendizaje no se logra un porcentaje bueno de individuos bien clasificados. Las causas de eso son un débil número de sujetos y robustez.

El algoritmo a tomado como parámetros raíz la rotación intervertebral de la unión superior y el índice de torsión.

6.2 Resultados del PERCEPTRON (red de neurones)

A continuación se muestran los resultados del algoritmo PERCEPTRON.

- Autovalidación

Se ha conseguido un 100% de sujetos bien clasificados y la matriz de confusión es:

Matriz de confusión

a	**b**	**←clasificado como**
24	0	**a = Estable**
0	25	**b = Evolutiva**

Tabla 8. Matriz de confusión para la autovalidación con PERCEPTRON

$$Sensibilidad(\%) = \frac{25}{25+0} = 100\%$$

$$Especificidad(\%) = \frac{24}{24+0} = 100\%$$

- Leave one out

Con esta modalidad se ha obtenido un 81.63% de sujetos clasificados de forma correcta.

Matriz de confusión

a	**b**	**←clasificado como**
20	4	**a = Estable**
5	20	**b = Evolutiva**

Tabla 9. Matriz de confusión para leave one out con PERCEPTRON

$$Sensibilidad(\%) = \frac{20}{20+5} = 80\%$$

$$Especificidad(\%) = \frac{20}{20+4} = 83.3\%$$

- Validación con 10 pacientes

En total 80% de sujetos bien clasificados.

Matriz de confusion

a	**b**	**←clasificado como**
3	2	**a = Estable**
0	5	**b = Evolutiva**

Tabla 10. Matriz de confusión para validación con 10 pacientes con PERCEPTRON

$$Sensibilidad(\%) = \frac{5}{5+0} = 100\%$$

$$Especificidad(\%) = \frac{3}{3+2} = 60\%$$

6.2.1 CONCLUSIONES DE LOS RESULTADOS CON PERCEPTRON

Se han logrado muy buenos valores de sensibilidad y especificidad para las dos primeras técnicas. Sin embargo, aunque la sensibilidad es del 100% con la validación sobre 10 pacientes, la especificidad es sólo del 60%. El problema con las redes neuronales es que los parámetros usados por el algoritmo para obtener los resultados son desconocidos, de ahí que no podamos utilizar el PERCEPTRON para conocer cuáles son las variables que hacen que la escoliosis evolucione incluso habiendo obtenido una buena predicción.

6.3 Resultados RULEFIT (Ensemble learning)

Con RULEFIT hemos lanzado el algoritmo con dos modalidades diferentes. Una con la que hemos obtenido la predicción de las escoliosis que se agravarán. Y otra con la que se han extraido reglas de asociación de las variables en principio claves para la agravación de la escoliosis.

- RULEFIT con autovalidación

Se han conseguido 100% de individuos bien clasificados y la matriz de confusión:

Matriz de confusión

a	b	**←clasificado como**
24	0	**a = Estable**
0	25	**b = Evolutiva**

Tabla 11. Matriz de confusión para la autovalidación con RULEFIT

$$Sensibilidad(\%) = \frac{25}{25+0} = 100\%$$

$$Especificidad(\%) = \frac{24}{24+0} = 100\%$$

- RULEFIT con “leave one out”

En total 67.35 sujetos clasificados correctamente.

Matriz de confusión

a	b	**←clasificado como**
17	7	**a = Estable**
9	16	**b = Evolutiva**

Tabla 12. Matriz de confusión para leave one out con RULEFIT

$$Sensibilidad(\%) = \frac{16}{16+9} = 64\%$$

$$Especificidad(\%) = \frac{17}{17+7} = 70.83\%$$

- Validación sobre 10 pacientes

Matriz de confusión

a	**b**	**←clasificado como**
3	2	**a = Estable**
0	5	**b = Evolutiva**

Tabla 13. Matriz de confusión para la validación con 10 pacientes con RULEFIT

$$Sensibilidad(\%) = \frac{5}{5+0} = 100\%$$

$$Especificidad(\%) = \frac{3}{3+2} = 60\%$$

Se ha obtenido un 80% de sujetos bien clasificados.

- Utilización de RULEFIT para encontrar reglas de asociación

A continuación se muestran los parámetros discriminantes elegidos por RULEFIT y los valores límites que definen la regla por orden de importancia para el modelo construido con el método "autovalidación".

Regla 1

Incidencia pelviana < 56.353

Rotación intervertebral axial de la zona de unión superior > -2.65

Regla 2

Rotación intervertebral axial de la zona de unión superior > -2.7

Rotación vertebral axial de la zona apical < 6.85

Regla 3

Rotación vertebral axial de la zona apical > 2.9

Relación entre la torsión y la rotación intervertebral axial de la zona de unión superior < -0.65

Regla 4

Rotación intervertebral axial de la zona de unión inferior < 2.7

Rotación intervertebral axial de la zona de unión superior > -3.65

Regla 5

Rotación intervertebral axial de la zona de unión superior >-2.7

Relación entre la torsión y la rotación intervertebral axial de la zona de unión superior > -1.8

Regla 6

Torsión > 3.45

Rotación intervertebral axial de la zona de unión inferior > 1.35

La predicción del algoritmo es una combinación lineal de las reglas anteriores.

6.3.1 CONCLUSIONES DE LOS RESULTADOS DE RULEFIT

De la misma forma que con las redes neuronales, hemos obtenido una buena sensibilidad y especificidad para las dos primera técnicas pero con la validación sobre 10 pacientes la especificidad ha sido del 60%.

En cuanto a la asociación de parámetros, vemos que los dos parámetros más frecuentes son las rotaciones intervertebrales axiales en las zonas de unión superior e inferior, siendo más discriminante la zona superior.

Estos resultados concuerdan con los obtenidos con los árboles de decisión por lo que a priori podemos decir que la rotación intervertebral en la zona superior es un parámetro que tiene fuerte influencia sobre la agravación de la escoliosis.

6.4 Resultados K-MEANS (clustering)

A continuación los resultados obtenidos con el algoritmo K-means.

- Autovalidación

	Cluster 1	Cluster 2
Estables	10	14
Evolutivos	11	14

Tabla 14. Matriz de confusión para la autovalidación con clustering

Hemos obtenido un 48.98% de sujetos mal clasificados.

6.4.1 CONCLUSIONES DE LOS RESULTADOS DEL K-MEANS

Podemos ver que el algoritmo ha clasificado la misma proporción de escoliosis evolutivas y de estables en los dos clusters. Eso quiere decir que si cogiéramos al hazar un sujeto del conjunto de datos, tendríamos la misma probabilidad de que fuera evolutivo que de que fuera estable por lo que este método no aporta nada para la predicción de la agravación de la escoliosis.

Si no se consiguen buenos resultados con la autovalidación es seguro que no los obtendremos tampoco con las otras técnicas por lo que paramos aquí el análisis con K-means.

6.5 Resultados con Q-finder

Estos son los resultados logrados con el algoritmo Q-finder. Se muestran las reglas con los parámetros discriminantes elegidos por el algoritmo y los valores límites que definen la regla. Estas cuatro reglas cubren la totalidad de las escoliosis evolutivas de la base de datos.

Regla 1: 18 evolutivos

Torsión entre 2.8 et 9.9

Rotación vertebral axial de la zona apical: entre 4 et 15.9

Sin desigualdad de los miembros inferiores

Regla 2: 17 evolutivos

Rotación intervertebral axial de la zona de unión superior entre -7 et -2.7

Regla 3: 17 evolutivos

Torsión entre 3.4 et 9.9

Relación entre Torsión et rotación intervertebral axial de la zona de unión inferior entre -0.6 et 3.

Regla 4: 16 evolutivos

Rotación intervertebral axial de la zona de unión inferior entre 1.4 et 5

Ángulo de Cobb Frontal entre 12.7 et 25.3

6.5.1 CONCLUSIONES DE LOS RESULTADOS CON Q-FINDER

Q-finder ha vuelto a elegir las rotaciones intervertebrales de las uniones superior e inferior como parámetros discriminantes. La regla 4, por ejemplo, cubre el % de todas las escoliosis evolutivas de la base. Muestra que si tenemos un ángulo de Cobb entre 12.7 y 25.3 y una rotación intervertebral axial de la zona de unión inferior entre 1.4 y 5 no hay ninguna escoliosis que permanecerá estable.

Con este algoritmo no podemos hacer la predicción por lo que no podemos hablar de sensibilidad o especificidad. El algoritmo a entregado 4 reglas que cubren la totalidad de las escoliosis evolutivas de la base de datos.

CAPÍTULO 7.CONCLUSIONES FINALES

Puesto que el objetivo del proyecto es encontrar métodos con los que poder predecir bien la evolución de la escoliosis en un primer examen médico, podemos afirmar que los resultados obtenidos con los diferentes algoritmos son buenos, sin embargo no podemos seleccionar un algoritmo predominante sobre los demás ya que cada algoritmo tiene sus ventajas e inconveniente que deben ser tenidos en cuenta en el momento de la selección y en función de las necesidades.

Preparación de los datos: Por otro lado, el funcionamiento de un método estadístico depende fuertemente del tratamiento de los datos y por tanto aunque el algoritmo elegido sea el correcto, sin la preparación previa de los datos correcta, los resultados obtenidos son serán buenos

Árboles de decisión: En lo que concierne a los árboles de decisión hemos obtenido muy buenos resultados trabajando con la autovalidación (mismo conjunto de datos para el aprendizaje y para la validación). Sin embargo, el porcentaje de individuos bien clasificados disminuye cuando utilizamos las otras técnicas. Una ventaja importante de los árboles de decisión es la posibilidad de determinar los parámetros más discriminantes puesto que uno de los objetivos iniciales del proyecto es saber cuáles son las variables que están ligadas a la agravación de la escoliosis.

Redes neuronales: Una red neuronal utilizando un algoritmo potente podría darnos resultados competentes. Sin embargo el mayor inconveniente de este método este que sólo podemos extraer como resultado el porcentaje de sujetos bien clasificados pero no sabemos cuáles son los parámetros utilizados para calcularlo. Este algoritmo es por tanto muy difícil de interpretar y de extraer informaciones que caractericen la agravación de la escoliosis.

K-means: En cuanto al K-means, no hemos obtenido resultados competentes ni para la predicción ni para la clasificación.

RULEFIT: Con RULEFIT hemos logrado una predicción débil con "leave one out", al contrario que la validación sobre 10 pacientes, que ha sido bastante buena. Inicialmente se había pensado en la utilización de RULEFIT como un algoritmo para encontrar cuáles son las variables con influencia en la agravación de la escoliosis, pero los resultados de la predicción demuestran que también es competente para esta aplicación.

Q-finder: En lo que concierne al Q-finder, podemos decir que es un algoritmo muy potente para encontrar relaciones entre los parámetros y el fenómeno a explicar. A la luz de los resultados, este algoritmo es el más adecuado para el objetivo y se presenta como una herramienta eficaz.

Parámetros importantes: Otro parte del objetivo del proyecto es conocer cuáles son los parámetros que hacen que la escoliosis se agrave. Podemos decir que los parámetros más frecuentes son la rotación intervertebral de la zona de unión superior, el índice de torsión y la rotación intervertebral de la zona de

unión inferior. Por lo que se puede afirmar que son variables de gran influencia sobre la agravación de la escoliosis.

Base de datos: Hemos trabajado con una base de datos de 49 pacientes. Después de analizar los algoritmos encontramos que el número de pacientes de bajo para obtener resultados competentes.

Elección de los algoritmos: Con el fin de hacer una buena selección de técnicas de "data mining" hemos hecho una fuerte investigación bibliográfica. Sin embargo, para el uso de los algoritmos dentro de cada método estadístico nos hemos basado en la opinión de expertos. Cada aplicación tiene sus necesidades y cada algoritmo tiene unas características diferentes. Para el futuro, sería aconsejable hacer una investigación profunda sobre los algoritmos disponibles para cada técnica con el fin de poder utilizar uno que se adapte mejor a las necesidades.

Futuro: Este trabajo aporta un bagaje sobre las familias de métodos existente y como pueden ser utilizados para identificar las escoliosis evolutivas.

Una vez tenemos información sobre esto, para el futuro, sería aconsejable estudiar más en detalles uno o dos de los métodos para poder obtener el mayor rendimiento posible. Además, habría que completar la base de datos actual con información sobre más pacientes.

BIBLIOGRAFÍA

1. "Comprendre et utiliser les statistiques dans les sciences de la vie" B. Falissard. 2005

2. "Data Mining et statistique décisionnelle. L'intelligence des données" TUFFERY Stéphane. 2010

3. "Redes Neuronales Artificiales. Fundamentos, Modelos y Aplicaciones". José Ramón Hilera González

4. M. Remzi, B. Djavan. Service d'urologie, Université médicale de Vienne, Währinger Gürtel 18-20, 1090 Vienne, Autriche

5 "Predictive learning via Rule Ensembles" . Friedman, J. H and Popescu, B. E. (Feb. 2005)

6. "Ensemble learning" Martin Sewel. Department of Computer Science University College London. August 2008

7. "Escoliosis: Realidad tridimensional" Miguel Ángel González Viejo, Oriol Cohí Riambau, Felip Salinas Castro. Ed. Masson

8. "Fisioterapia para la escoliosis basada en el diagnóstico". Hans Rudolf Weiss, Manuel Rigo. Ed. Paidotribo

9. "Aplicación de la minería de datos al estudio de las alteraciones respiratorias durante el sueño" Carlos Zamarrón Sanz, Vanesa García Paz, Uxío Calvo Álvarez, Fernanda Pichel Guerrero, José Ramón Rodríguez Suáre. Servicio de Neumología. Hospital Clínico Universitario de Santiago de Compostel.

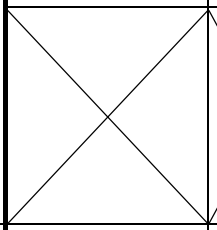
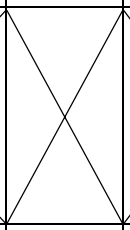
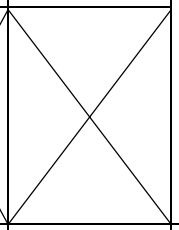
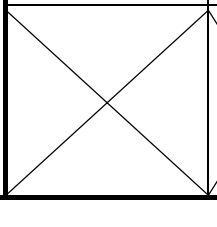
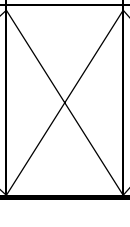
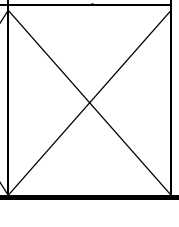
10."Análisis de información clínica mediante técnivas de minería de datos" Ingrid Wilford Rivera (Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría, Cuba), Alejandro Rosete Suárez (Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría, Cuba) Alfredo Rodríguez Díaz (Centro para el Desarrollo Informático en la Salud, MINSAP, Cuba)

11. "Pruebas diagnósticas: Sensibilidad y especificida" Pita Fernández, S., Pértegas Díaz, S. Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario-Universitario Juan Canalejo. A Coruña (España).

12. "Recherche des facteurs biomécaniques dans l'aggravation des scolioses idiopathiques" Nicolas CHAMPAIN. 2004

13"Les scolioses" J.J. Rainaut. Ed. Ellipses

14. "Anatomía con orientación clínica"Cuarta edición. Keith L. Moore, Arthur F. Dalley. Ed.Panamericana

TÉCNICA	ALGORITMOS	ALGORITMO UTILIZADO	VENTAJAS	INCONVENIENTES	PREDICCIÓN: 3 TÉCNICAS			PREDICCIÓN	EXTRACCIÓN DE INFORMACIÓN
Árboles de decisión	Random forest Random multinomial logit C4.5 CART	C4.5	Muestran claramente cuales son los parámetros discriminantes	Los nodos del nivel n+1 dependen fuertemente de los del nivel n. Capacidad de predicción débil	**Autovalidación** 95.92%	**Leave one out** 67.35%	**Validación 10** 40%	–	+
Clustering	Anes Mona Diana Cobweb K-means Pam	K-MEANS	Podemos manejar sólo los clusters en lugar de todo el conjunto de datos Los clusters son definidos a partir del mismo conjunto de datos	Depende mucho del número de cluster elegido Capacidad de predicción nula	**Autovalidación** 	**Leave one out** 	**Validación 10** 	– – –	–
Redes neuronales	Kohonen map Hopfield net Perceptron Adaline	PERCEPTRON	Herramienta muy eficaz para la predicción	Son necesarios muchas muestras para el aprendizaje No sabemos como el algoritmo utiliza las variables	**Autovalidación** 100%	**Leave one out** 81.63%	**Validación 10** 80%	+ + +	– – –
Ensemble learning	Bagging Boosting AdaBoost Bayes Optimal Classifier	RULEFIT	Pueden ser utilizados para la predicción y para la clasificación Son fáciles de interpretar		**Autovalidación** 100%	**Leave one out** 67.35%	**Validación 10** 80%	+	+
Règles d'association	Q-FINDER	Q-FINDER	Se puede modificar las reglas manualmente para encontrar nueva información	Nopodemos hacer la predicción	Autovalidation 	Leave 	Validation 	– – –	+ + +